Programovanie v prostredí

IV. Základy štatistiky

Aleš Kozubík

Katedra Matematických Metód a Operačnej Analýzy

25.11.2019

Používané dáta

Budeme používať datasety `iris`, `trees`, `warpbreaks` a `PlantGrowth`, ktoré sú obsiahnuté v samotnom systéme R.

S ich obsahom sa môžeme bližšie oboznámiť pomocou funkcie `help(<meno>)`, kde `<meno>` je názov jednotlivých datasetov.

Čo si ukážeme

- Výpočet výberových charakteristík
- Miery asociácie dát (štatistickej závislosti)
- Štatistické testy.

Sumárny prehľad výberových štatistík

Prehľadný súhrn hlavných štatistík získame pomocou funkcie `summary()`

```
> summary(iris)
 Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
Median :5.800   Median :3.000   Median :4.350   Median :1.300
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Prehľad funkcií pre výberové štatistiky

Z dispozícii máme tieto funkcie:

Štatistika	**Funkcia**	**Štatistika**	**Funkcia**
Stredná hodnota	`mean`	Medián	`median`
Smerodajná odchýlka	`sd`	Rozptyl	`var`
Maximum	`max`	Minimum	`min`
Medzikvartilové rozpätie	`IQR`	Rozpätie	`range`
Súčet	`sum`	Kvantily	`quantiles`
Počet meraní	`length`	Súčin	`prod`
Tukeyova súhrnná päťčíselná charakteristika			`fivenum`

Výberová stredná hodnota (priemer)

Vypočítame pomocou funkcie `mean()`. Jej argumentom je vektor numerických hodnôt, z ktorých chceme priemer počítať.

Ak vektor obsahuje nezistené hodnoty (implementované ako hodnota `NA`), tak pomocou parametra `na.rm=T` zariadime ich vynechanie z výpočtov. Inak výpočet zlyhá, výsledkom je `NA`.

Funkcia `sapply()`

Ak chceme aplikovať priemer resp. inú funkciu na všetky skupiny v rámci určitej množiny dát (datasetu), použijeme funkciu `sapply()`, ktorej argumentmi budú názov datasetu a charakteristika, ktorú chceme určiť.

Napríklad:

```
> sapply(trees,mean)
   Girth   Height   Volume
13.24839 76.00000 30.17097
```

Funkcia sapply()

POZOR: Všetky údaje musia byť numerické. Ak premenná obsahuje nenumerické hodnoty – treba ich z výpočtu vynechať

Napríklad

```
sapply(iris, max)
Error in Summary.factor(c(1L, 1L, 1L, 1L, 1L, :
  max not meaningful for factors
```

Preto:

```
> sapply(iris[-5], max)
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
         7.9          4.4          6.9          2.5
```

Štatistiky podľa skupín

Často chceme zhrnúť určité numerické hodnoty podľa úrovne nejakého faktora.

Vtedy použijeme funkciu `tapply()` s niektorou zo štatistík uvedených v tabuľke.

```
> tapply(iris$Sepal.Width,iris$Species,mean)
    setosa versicolor  virginica
     3.428      2.770      2.974
```

Štatistiky podľa skupín

Inou alternatívou je použitie funkcie `aggregate()`

Výhodou je možnosť simultánne sumarizovať viaceré spojité premenné a taktiež možnosť použiť štatistické funkcie, ktoré dávajú ako výsledok viac hodnôt (ako `range()` alebo `quantile()`.

```
> aggregate(Sepal.Width~Species, iris, mean)
     Species Sepal.Width
1     setosa       3.428
2 versicolor       2.770
3  virginica       2.974
```

Štatistiky podľa skupín

Sumarizácia viacerých hodnôt

Ak chceme simultánne sumarizovať charakteristiky viacerých spojitých premenných, zlúčime ich pomocou funkcie `cbind()`.

```
> aggregate(cbind(Sepal.Width,Sepal.Length)~Species,
+ iris,mean)
     Species Sepal.Width Sepal.Length
1     setosa       3.428        5.006
2 versicolor       2.770        5.936
3  virginica       2.974        6.588
```

Miery asociácie (štatistickej závislosti)

Kovariancia je mierou lineárnej závislosti dvoch spojitých premenných. Je závislá od jednotiek merania.
Ak máme náhodný výber dvojíc $(x_1, y_1), \ldots, (x_n, y_n)$, definujeme

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

a kovarianciu S_{xy} ako

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Miery asociácie (štatistickej závislosti)

Pearsonov korelačný koeficient je mierou závislosti nezávislou od jednotiek merania. Definujeme

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2, \quad S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \overline{y})^2$$

Potom

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}}.$$

Miery asociácie (štatistickej závislosti)

Spearmanov poradový korelačný koeficient je neparametrickou alternatívou k Pearsonovmu koeficientu, meria lineárnu aj nelineárnu závislosť.

Ak máme náhodný výber dvojíc $(x_1, y_1), \ldots, (x_n, y_n)$, a nech $R_1, \ldots, R_n$ sú poradia hodnôt $x_1, \ldots, x_n$ a $S_1, \ldots, S_n$ poradia hodnôt $y_1, \ldots, y_n$ v usporiadaných výberoch.

Spearmanov poradový korelačný koeficient sa potom definuje ako korelačný koeficient dvojíc $(R_1, S_1), \ldots, (R_n, S_n)$.

Výpočet kovariancie a korelácie

Kovariancia a korelácia dvoch premenných

Pre výpočet kovariancie používame funkciu `cov()` a pre výpočet korelácie funkciu `cor()`.

Nech náhodný výber z dvoch premenných je v podobe vektorov `x` a `y`. Potom použijeme `cov(x,y,)` resp. `cor(x,y)`.

```
> cov(trees$Height,trees$Volume)
[1] 62.66
> cor(trees$Height,trees$Volume)
[1] 0.5982497
```

Výpočet kovariancie a korelácie

Kovariančná a korelačná matica

Pomocou funkcií `cov()` a `cor()` môžeme vypočítať aj celú kovariančnú resp. korelačnú maticu pre celú množinu dát.

```
> cov(trees)
             Girth    Height     Volume
Girth    9.847914 10.38333   49.88812
Height  10.383333 40.60000   62.66000
Volume  49.888118 62.66000 270.20280
> cor(trees)
             Girth     Height     Volume
Girth   1.0000000 0.5192801 0.9671194
Height  0.5192801 1.0000000 0.5982497
Volume  0.9671194 0.5982497 1.0000000
```

Výpočet kovariancie a korelácie

Kovariančná a korelačná matica – poznámky

Ak počítame kovarianciu resp. koreláciu dvoch premenných, obidva vektory musia mať rovnakú dĺžku.

Ak sa v dátovej množine vyskytujú neznáme hodnoty, je potrebné ako argument `use` funkcie `cov()` resp. `cor()` nastaviť hodnotu `"pairwise"`. Inak by kovariančná matica obsahovala taktiež neurčené hodnoty.

```
> x<-c(1,2,3,4,5)
> y<-c(1,2,3)
> y[5]<-5
> cov(x,y,use="pairwise")
[1] 2.916667
```

Výpočet kovariancie a korelácie

Kovariančná a korelačná matica – poznámky

Všetky hodnoty, ktoré vstupujú do výpočtu musia byť numerické, inak dôjde ku chybe.

```
> cov(iris)
Error: is.numeric(x) || is.logical(x) is not TRUE
> cov(iris[-c(4,5)])
              Sepal.Length Sepal.Width Petal.Length
Sepal.Length    0.6856935  -0.0424340    1.2743154
Sepal.Width    -0.0424340   0.1899794   -0.3296564
Petal.Length    1.2743154  -0.3296564    3.1162779
```

Výpočet kovariancie a korelácie

Spearmanov koeficient poradovej korelácie

Pre výpočet Spearmanovho koeficientu poradovej korelácie môžeme opäť použiť funkciu `cor()`, je však potrebné nastaviť parameter `method` na hodnotu `"spearman"`

```
> cor(trees$Height,trees$Volume,method="spearman")
[1] 0.5787101
```

Interval spoľahlivosti

Všeobecný pojem intervalu spoľahlivosti

Interval spoľahlivosti vo všeobecnosti charakterizujeme ako interval, do ktorého padnú hodnoty určitej štatistiky s danou pravdepodobnosťou.

Využívame ho ako intervalový odhad, t.j. na základe náhodného výberu zostrojíme interval, v ktorom sa hľadaná hodnota parametra nachádza s požadovanou pravdepodobnosťou.

Intervaly spoľahlivosti

Podiel populácie (ukazovateľ štruktúry súboru)

Chceme určiť podiel p prvkov v danej populácii (pokusnom súbore), ktorý má sledovanú vlastnosť, pričom máme k dispozícii náhodný výber o rozsahu n, na základe ktorého sme odhadli podiel $\hat{p}$. Náhodná premenná

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

má štandardizované normálne rozdelenie $N(0,1)$.
$(1-\alpha)\cdot 100\%$ interval spoľahlivosti má tvar:

$$\left(\hat{p} - q_{\frac{\alpha}{2}} \cdot \sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + q_{\frac{\alpha}{2}} \cdot \sqrt{\hat{p}(1-\hat{p})/n}\right),$$

kde $q_{\frac{\alpha}{2}}$ je tzv. kritická hodnota rozdelenia $N(0,1)$.

Intervaly spoľahlivosti

Podiel populácie – príklad

V telefonickom prieskume odpovedalo 1 013 respondentov na otázku, ako sú spokojní s výkonom funkcie prezidenta. 466 respondentov uviedlo dobre alebo výborne. Aký je 95% interval spoľahlivosti pre túto hodnotu?

```
> n<-1013
> odhad<-466/n
> sd<-sqrt(odhad*(1-odhad)/n)
> alfa<-0.05
> q<-qnorm(alfa/2)
> q
[1] 1.959964
> c(odhad-q*sd,odhad+q*sd)
[1] 0.4293281 0.4907114
```

Intervaly spoľahlivosti

Podiel populácie – príklad

Úlohu by bolo možné riešiť aj pomocou funkcie `prop.test()` alebo `binom.test()`, ktorá namiesto normálneho rozdelenia používa adekvátne binomické rozdelenie.

```
> prop.test(466,1013,conf.level=0.95)

1-sample proportions test with continuity correction

data:  466 out of 1013, null probability 0.5
X-squared = 6.3179, df = 1, p-value = 0.01195
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4290475 0.4912989
sample estimates:
      p
```

Intervaly spoľahlivosti

Stredná hodnota

Ak $X_1, X_2, \ldots, X_n$ je náhodný výber zo súboru so strednou hodnotou μ a smerodajnou odchýlkou σ, tak $(1-\alpha)\cdot 100\%$ interval spoľahlivosti je

$$\overline{X} \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}},$$

kde $t_{\frac{\alpha}{2}}$ je kritická hodnota Studentovho rozdelenia a s je výberová smerodajná odchýlka.

Intervaly spoľahlivosti

Stredná hodnota – príklad

V triede je 30 študentov, ich priemerná výška je 180 cm a smerodajná odchýlka výšky je 10 cm. Aký je 80 % interval spoľahlivosti pre ich výšku?

```
> n<-30
> priemer<-180
> sd<-10
> alfa<-0.2
> t<-qt(1-alfa/2,df=n-1)
> se<-sd/sqrt(n)
> c(priemer-t*se,priemer+t*se)
[1] 177.6057 182.3943
```

Intervaly spoľahlivosti

Smerodajná odchýlka

Nech $X_1, X_2, \ldots, X_n$ je náhodný výber, l a r sú kritické hodnoty rozdelenia χ^2, také, že

$$\mathbb{P}(l \leq \chi^2_{n-1} \leq r) = 1 - \alpha.$$

Potom interval spoľahlivosti pre smerodajnú odchýlku σ je

$$\left(\frac{(n-1)s^2}{r}, \frac{(n-1)s^2}{l} \right).$$

Intervaly spoľahlivosti

Smerodajná odchýlka – výpočet

Pre výpočty použijeme kvantilovú funkciu `qchisq()` rozdelenia chí-kvadrát.

Ktirické hodnoty teda sú

```
> l<-qchisq(alfa/2,df=n-1)
> r<-qchisq(1-alfa/2,df=n-1)
```

Intervaly spoľahlivosti

Smerodajná odchýlka – výpočet

Príklad

Doba jazdy medzi dvomi zastávkami sa riadi normálnym rozdelením s neznámou strednou hodnotou a rozptylom. Pre lepšiu predstavu o variabilite časov chceme odhadnúť smerodajnú odchýlku.
Na základe 100 meraní bola zistená priemerná doba jazdy 20 minút s rozptylom rovným 10.

Intervaly spoľahlivosti

Smerodajná odchýlka – riešenie

```
> s2<-10; n<-100
> alfa<-0.05
> lavy<-qchisq(alfa/2,df=n-1)
> pravy<-qchisq(1-alfa/2,df=n-1)
# pre rozptyl
> (n-1)*s2*c(1/pravy,1/lavy)
[1]  7.70896 13.49489
# pre smerodajnu odchylku
> sqrt((n-1)*s2*c(1/pravy,1/lavy))
[1] 2.776501 3.673540
```

Podstata testov významnosti

Overuje sa platnosť tzv. nulovej hypotézy H_0 oproti alternatívnej hypotéze H_A. Obvykle ide o tvrdenie o hodnote nejakého parametra.

Na základe náhodného výberu (experimentu) sa zostrojí tzv. testovacia štatistika.

Na základe príslušnosti vypočítanej hodnoty testovacej štatistiky do intervalu spoľahlivosti zodpovedajúceho požadovanej hladine významnosti buď zamietneme alebo nezamietneme nulovú hypotézu.

Typy testov významnosti

Existujú tri druhy testov:

1. **Jednovýberový test** na základe náhodného výberu sa overuje hypotéza o rovnosti parametra predpokladanej hodnote.
2. **Dvojvýberový test** na základe dvoch nezávislých náhodných výberov sa testuje zhoda parametrov, t.j. či pochádzajú zo súborov s rovnakými charakteristikami.
3. **Párový test** medzi jednotlivými hodnotami vo dvoch náhodných výberoch je jednoznačná korešpondecia, musia mať rovnaký rozsah.

Test významnosti pre podiel populácie

Nulová hypotéza $H_0 : p = p_0$, tj. podiel p_0 zo základného súboru má požadovanú vlastnosť.

Alternatívna hypotéza je $H_A : p \neq p_0$ pre obojstrannú alternatívu, resp. $H_A : p > p_0$ alebo $H_A : p < p_0$ pre jednostrannú alternatívu.

Ak $\hat{p}$ je odhad podielu p, tak testovacia štatistika má tvar

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}.$$

Ak je hypotéza pravdivá, má Z štandardné normálne rozdelenie $N(0, 1)$.

Test významnosti pre podiel populácie

Normálne rozdelenie – graf

Používame funkciu `prop.test()`. Má niekoľko bežných argumentov:

- `x` počet prvkov s danou vlastnosťou vo výbere,
- `n` rozsah náhodného výberu,
- `p` predpokladaný podiel prvkov s danou vlastnosťou,
- `conf.level` hladina spoľahlivosti (default je 0.95),
- `alt` či ide o obojstranný test (alternatívna hypotéza $H_A : p \neq p_0$) alebo jednostranný test (alternatívna hypotéza $H_A : p > p_0$ hodnota `greater`, alebo $H_A : p < p_0$ hodnota `less`)

Test významnosti pre podiel populácie

príklad

V skupine 1600 osôb bolo objavených 850 osôb s nadváhou. Overte platnosť hypotézy, že nadpolovičná časť populácie má nadváhu.
Riešenie:

```
> prop.test(x=850,n=1600,p=0.5,alt="greater")
   1-sample proportions test with continuity correction

data:  850 out of 1600, null probability 0.5
X-squared = 6.1256, df = 1, p-value = 0.006662
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.5103813 1.0000000
sample estimates:
      p
0.53125
```

Test významnosti pre strednú hodnotu

Jednovýberový test hypotézy $H_0 : \mu = \mu_0$ proti alternatíve $H_A : \mu \neq \mu_0$ resp. jednostranné alternatívy $H_A : \mu > \mu_0, H_: \mu < \mu_0$

Testovacia štatistika má tvar

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}}.$$

Ak je hypotéza pravdivá, tak T má Studentovo rozdelenie $t(n-1)$ s $n-1$ stupňami voľnosti.

Test významnosti pre strednú hodnotu

Na vykonanie testu slúži funkcia `t.test()` s argumentmi:

- `x` premenná obsahujúca prvky náhodného výberu,
- `mu` predpokladaná stredná hodnota,
- `conf.level` hladina spoľahlivosti (default je 0.95),
- `alt` či ide o obojstranný alebo jednostranný test (default je obojstranný).

Test významnosti pre strednú hodnotu

Príklad

S testovaným vozidlom bolo vykonaných 10 jázd a postupne nameraná priemerná spotreba 10.1, 12, 9.6, 10.5, 11, 9.8, 10.6, 10.7, 9.7, 11 l/100 km. Chceme overiť, či výsledky zodpovedajú deklarovanej spotrebe 10 l km

```
> x<-c(10.1,12,9.6,10.5,11,9.8,10.6,10.7,9.7,11)
> t.test(x,mu=10,conf.level=0.99,alt="greater")

        One Sample t-test

data:  x
t = 2.1429, df = 9, p-value = 0.03037
alternative hypothesis: true mean is greater than 10
99 percent confidence interval:
 9.841664      Inf
sample estimates:
mean of x
     10.5
```